

Quantum Jarzynski-Sagawa-Ueda relations

Yohei Morikuni¹ and Hal Tasaki¹

Abstract

We consider a (small) quantum mechanical system which is operated by an external agent, who changes the Hamiltonian of the system according to a fixed scenario. In particular we assume that the agent (who may be called a demon) performs measurement followed by feedback, i.e., it makes a measurement of the system and changes the protocol according to the outcome. We extend to this setting the generalized Jarzynski relations, recently derived by Sagawa and Ueda for classical systems with feedback. One of the two relations by Sagawa and Ueda is derived here in error-free quantum processes, while the other is derived only when the measurement process involves classical errors. The first relation leads to a second law which takes into account the efficiency of the feedback.

1 Introduction

Recent progress in statistical physics has led to nontrivial exact relations such as the fluctuation theorem [1, 2, 3] and the Jarzynski relation [4, 3], which hold even when physical systems are driven out of equilibrium. These relations reveal rich structures hidden in the canonical formalism of equilibrium states, and also suggest that there can be some universal structures out of equilibrium. Since these exact relations mainly deal with fluctuation, they are expected to find useful applications in small systems, where relevant energy scale is comparable to that of thermal fluctuation.

In recent papers [5, 6], Sagawa and Ueda studied the effect of feedback or the Maxwell demon in (small) thermodynamic systems. They showed that the second law of thermodynamics and the Jarzynski relation should be properly modified in order to take into account the information of the system gained by the demon. In a recent beautiful experiment, Toyabe, Sagawa, Ueda, Muneyuki, and Sano [7] have designed a system with an artificial demon intervening the time evolution, and clearly demonstrated that one can convert information into useful work². They also confirmed the validity of one of the generalized Jarzynski relations obtained by Sagawa and Ueda [6], which we shall call Jarzynski-Sagawa-Ueda relations.

In the present paper, we treat a quantum system with feedback, and look for extensions of the Jarzynski-Sagawa-Ueda relations, which were proved only for classical systems. There may be a possibility that such extensions will become relevant to experiments or manipulations of small quantum systems. But let us stress that by studying quantum extensions one will get a better understanding of basic structures of the relations, e.g., which relation is universal and which is intrinsic to classical systems. We

¹ Department of Physics, Gakushuin University, Mejiro, Toshima-ku, Tokyo 171-8588, Japan.

² The term “information-to-energy conversion” found in the title of [7] should better be read “information-to-free-energy conversion.”

believe that such an attempt is of interest from a purely theoretical point of view. In fact we found that one of the two relations derived by Sagawa and Ueda can be derived in an error-free quantum process, while the other can be (for the moment) shown only when we include somewhat artificial classical errors into the process. For other related works on processes including feedback and/or measurement, see [8, 9] and references therein.

The present paper is organized as follows. In section 2, we study the most basic setting with no errors and derive the quantum extension of one of the two relations by Sagawa and Ueda. In section 3, we introduce classical errors to the feedback process, and derive a quantum counterpart of the other relation. In section 4, we discuss some variations and extensions of the main results as well as a related interesting issue of “clockwork demon.”

2 Error-free feedback

Let us start from an ideal setting with no errors, and present our main result.

Setting We consider an isolated quantum mechanical system³ whose Hamiltonian has parameters which can be controlled by an outside agent. The system is initially prepared in the equilibrium state. In the first stage of the time-evolution, the Hamiltonian is changed according to a fixed protocol. Then one makes a measurement on the system. The second stage of the time evolution takes into account feedback from the measurement, and the Hamiltonian is changed according to a protocol which depends on the outcome of the measurement.

Let us be more precise. By $H^{(0)}$ we denote the initial Hamiltonian of the system. Its normalized eigenstates and the corresponding eigenvalues are denoted as φ_i and $E_i^{(0)}$, respectively, with $i = 1, 2, \dots$. The projection onto φ_i is denoted as $P_i^{(0)}$. We write the corresponding canonical density matrix as⁴

$$\rho_{\text{can}}^{(0)} := \frac{e^{-\beta H^{(0)}}}{Z_0} = \sum_i \frac{e^{-\beta E_i^{(0)}}}{Z_0} P_i^{(0)}, \quad (2.1)$$

with $Z_0 := \sum_i e^{-\beta E_i^{(0)}}$. Finally $F^{(0)} := -\beta^{-1} \log Z_0$ is the corresponding free energy.

Since we perform feedback, the Hamiltonian $H^{(j)}$ at the final moment depends on the outcome $j = 1, \dots, n$ of the intermediate measurement. We denote by $\psi_k^{(j)}$ and $E_k^{(j)}$ the normalized eigenstate and the corresponding eigenvalue, respectively, of $H^{(j)}$ with $k = 1, 2, \dots$. The projection onto $\psi_k^{(j)}$ is denoted as $P_k^{(j)}$. Again we write the

³ The system can be a small (and literally isolated) one, or a combination of a (small) system of interest and a larger system which plays the role of the heat bath.

⁴ By $A := B$ or $B =: A$, we mean that A is defined in terms of B .

corresponding canonical density matrix as

$$\rho_{\text{can}}^{(j)} := \frac{e^{-\beta H^{(j)}}}{Z_j} = \sum_k \frac{e^{-\beta E_k^{(j)}}}{Z_j} P_k^{(j)}, \quad (2.2)$$

with $Z_j := \sum_k e^{-\beta E_k^{(j)}}$, and the free energy as $F^{(j)} := -\beta^{-1} \log Z_j$.

We consider the following process in the line of [10] (see also [11]). Initially the system is in the equilibrium state $\rho_{\text{can}}^{(0)}$. At the initial moment, one makes a projective measurement of the energy $H^{(0)}$, whose outcome is denoted as E_i or simply⁵ i . Then the Hamiltonian is changed according to a fixed protocol for a certain amount of time, and the state of the system evolves by an unitary operator U . This is the first stage. Then one makes a projective measurement⁶ with outcomes $j = 1, \dots, n$. We assume that the measurement is described by a set of projection operators Π_1, \dots, Π_n such that $\sum_{j=1}^n \Pi_j = 1$ and $\Pi_j \Pi_{j'} = 0$ if $j \neq j'$. The rest of the time evolution, which is the second stage, depends on the outcome j . The Hamiltonian is changed according to a fixed protocol associated with j , and the state evolves by an unitary operator U_j . Finally one makes a projective measurement of the final Hamiltonian $H^{(j)}$, whose outcome is denoted as $E_k^{(j)}$ or k .

The probability that one gets successive outcomes i, j, k in the above process is given by

$$p(i \rightarrow j \rightarrow k) := \text{Tr}[P_k^{(j)} U_j \Pi_j U P_i^{(0)} U^\dagger \Pi_j U_j^\dagger P_k^{(j)}] \frac{e^{-\beta E_i^{(0)}}}{Z_0}. \quad (2.3)$$

One can easily verify that it is normalized as $\sum_{i,j,k} p(i \rightarrow j \rightarrow k) = 1$. We define the average with respect to $p(i \rightarrow j \rightarrow k)$ as

$$\langle f(i, j, k) \rangle_p := \sum_{i,j,k} f(i, j, k) p(i \rightarrow j \rightarrow k), \quad (2.4)$$

where $f(i, j, k)$ is an arbitrary function of i, j , and k .

Main result In this setting we show that

$$\left\langle \exp \left[\beta \{ W_{i,j,k} - (F^{(0)} - F^{(j)}) \} \right] \right\rangle_p = \gamma, \quad (2.5)$$

where $W_{i,j,k} := E_i^{(0)} - E_k^{(j)}$, and

$$\gamma := \sum_j \text{Tr}[\Pi_j U_j^\dagger \rho_{\text{can}}^{(j)} U_j \Pi_j]. \quad (2.6)$$

⁵ For simplicity we assume that $H^{(0)}$ has no degeneracy so that an accurate measurement of the energy uniquely determines i . But this assumption is not essential.

⁶ It is trivial to extend the results to the case where one makes measurement and feedback repeatedly as in [9].

This is the quantum extension of one of the generalized Jarzynski relations derived by Sagawa and Ueda [6]. The equality (2.5) corresponds to equation (6) in [6].

Since $W_{i,j,k}$ is the difference between the initial and the final energy of the system, it is natural to identify it with the work done by the system. But let us remark that this definition relies on the rather artificial setting where one precisely measures the energy in the initial and the final moments (but see the beginning of section 4, where we discuss the corresponding inequality). We also note that, in quantum systems, exchange of work also takes place during measurement processes.

The quantity $\text{Tr}[\Pi_j U_j^\dagger \rho_{\text{can}}^{(j)} U_j \Pi_j]$ which appears in (2.6) can be interpreted as the probability that one observes Π_j when the system starts from the equilibrium state $\rho_{\text{can}}^{(j)} = e^{-\beta H^{(j)}}/Z_j$ and evolves by the inverse time-evolution U_j^\dagger . This probability is expected to be close to one if the time-evolution U_j is chosen in such a way that any state within the range of Π_j evolves into a state not too far from the equilibrium state $\rho_{\text{can}}^{(j)}$. In such a case, γ becomes much larger than unity, suggesting that the feedback is designed in an efficient manner. Note that γ can be less than unity for a badly designed feedback. See the example below. As is stressed by Sagawa and Ueda [6], a remarkable feature of the quantity γ is that it can be measured by independent experiments.

Derivation By noting that $\sum_i P_i^{(0)} = 1$, and using the property of the trace, one gets

$$\begin{aligned} \left\langle e^{\beta(E_i^{(0)} - E_k^{(j)})} \frac{Z_0}{Z_j} \right\rangle_p &= \sum_{i,j,k} \text{Tr}[P_k^{(j)} U_j \Pi_j U P_i^{(0)} U^\dagger \Pi_j U_j^\dagger P_k^{(j)}] \frac{e^{-\beta E_k^{(j)}}}{Z_j} \\ &= \sum_{j,k} \text{Tr}[P_k^{(j)} U_j \Pi_j U_j^\dagger P_k^{(j)}] \frac{e^{-\beta E_k^{(j)}}}{Z_j} \\ &= \sum_{j,k} \text{Tr}[\Pi_j U_j^\dagger P_k^{(j)} U_j \Pi_j] \frac{e^{-\beta E_k^{(j)}}}{Z_j}. \end{aligned} \quad (2.7)$$

By recalling (2.2), we see that the right-hand side is equal to γ of (2.6). By noting that $Z_0/Z_j = \exp[-\beta(F^{(0)} - F^{(j)})]$, we get the desired relation (2.5).

Example As a simple illustrative example, consider a two-level system. We set $U = 1$ and $\Pi_j = P_j^{(0)}$, i.e., the intermediate measurement is the same as the initial measurement of the energy. As for the feedback, we set $U_1 = 1$ and let U_2 be the operator which simply switches the two eigenstates of $H^{(0)}$. We also set $H^{(1)} = H^{(2)} = H^{(0)}$. Then we have $p(1 \rightarrow 1 \rightarrow 1) = e^{-\beta E_1^{(0)}}/Z_0$ and $p(2 \rightarrow 2 \rightarrow 1) = e^{-\beta E_2^{(0)}}/Z_0$ with $Z_0 = e^{-\beta E_1^{(0)}} + e^{-\beta E_2^{(0)}}$, and $p(i \rightarrow j \rightarrow k) = 0$ for all other combinations. We then find

$$\left\langle e^{\beta(E_i^{(0)} - E_k^{(j)})} \right\rangle_p = \frac{2 e^{-\beta E_1^{(0)}}}{e^{-\beta E_1^{(0)}} + e^{-\beta E_2^{(0)}}}. \quad (2.8)$$

When $E_1^{(0)} < E_2^{(0)}$, the right-hand side, which is γ , is clearly larger than unity. This reflects the fact that the demon has made a clever use of the information to get extra

work from the system. The case $E_1^{(0)} > E_2^{(0)}$, where γ become less than unity, provides an example of a failed demon who made a wrong use of the information to lose work.

3 Feedback with classical errors

Next we consider a feedback process which includes errors, and extend the other relation obtained by Sagawa and Ueda.

Setting We consider almost the same process as in section 2, but assume that the intermediate measurement now involves errors. We assume that the errors are of purely classical nature⁷, i.e., when the intermediate measurement (described by Π_1, \dots, Π_n) yields the result j , one mis-interprets the result as j' with a given probability⁸ $\epsilon(j \rightarrow j') \geq 0$. The probability is normalized as $\sum_{j'} \epsilon(j \rightarrow j') = 1$ for any j . The error-free process considered in section 2 is recovered by setting $\epsilon(j \rightarrow j') = \delta_{j,j'}$. In what follows we assume that for each j' , the probability $\epsilon(j \rightarrow j')$ is either vanishing for all j or nonvanishing for all j .

The rest of the process (i.e., the second stage of the time evolution and the final measurement) is executed according to the result j' (not j). This means that the probability (2.3) is modified as

$$\tilde{p}(i \rightarrow j \rightarrow j' \rightarrow k) := \text{Tr}[P_k^{(j')} U_{j'} \Pi_j U P_i^{(0)} U^\dagger \Pi_{j'} U_{j'}^\dagger P_k^{(j')}] \epsilon(j \rightarrow j') \frac{e^{-\beta E_i^{(0)}}}{Z_0}. \quad (3.1)$$

As in (2.4), we denote the average over this probability as

$$\langle f(i, j, j', k) \rangle_{\tilde{p}} := \sum_{i, j, j', k} f(i, j, j', k) \tilde{p}(i \rightarrow j \rightarrow j' \rightarrow k). \quad (3.2)$$

Let us also define

$$\tilde{p}_2(j) := \sum_{i, j', k} \tilde{p}(i \rightarrow j \rightarrow j' \rightarrow k), \quad \tilde{p}_3(j') := \sum_{i, j, k} \tilde{p}(i \rightarrow j \rightarrow j' \rightarrow k), \quad (3.3)$$

and

$$\tilde{p}_{2,3}(j, j') := \sum_{i, k} \tilde{p}(i \rightarrow j \rightarrow j' \rightarrow k), \quad (3.4)$$

which are the probabilities to observe j , to observe j' , and to observe a pair (j, j') , respectively. As in [6], we define the (unaverage) mutual information as

$$I_{j, j'} := \log \frac{\epsilon(j \rightarrow j')}{\tilde{p}_3(j')} = \log \frac{\tilde{p}_{2,3}(j, j')}{\tilde{p}_2(j) p_3(j')}, \quad (3.5)$$

⁷ We admit that this setting is artificial. The main motivation for studying it is that we can derive the second Jarzynski-Sagawa-Ueda relation (3.7) only in this setting.

⁸ In the notation of [6], $\epsilon(j \rightarrow j')$ should read $P[j'|j]$.

where the second equality comes from $\tilde{p}_{2,3}(j, j') = \tilde{p}_2(j) \epsilon(j \rightarrow j')$. By averaging this, one gets

$$\langle I_{j,j'} \rangle_{\tilde{p}} = \sum_{j,j'} \tilde{p}_{2,3}(j, j') \log \frac{\tilde{p}_{2,3}(j, j')}{\tilde{p}_2(j) \tilde{p}_3(j')}, \quad (3.6)$$

which is the mutual information.

Main results In this setting we show that

$$\left\langle \exp \left[\beta \{ W_{i,j',k} - (F^{(0)} - F^{(j')}) \} - I_{j,j'} \right] \right\rangle_{\tilde{p}} = 1, \quad (3.7)$$

where the “work” is defined by $W_{i,j',k} := E_i^{(0)} - E_k^{(j')}$ as before. This is a quantum version of the other generalized Jarzynski relation, equation (4) in [6], derived by Sagawa and Ueda. Interestingly, this type of equality seems to be derivable only in the present setting with classical errors. Indeed the mutual information (3.5), (3.6) is a purely classical quantity as opposed to the QC-mutual information (see [5]), which takes into account the full quantum nature of the system⁹. See the discussion about the error-free limit at the end of the present section, and about the corresponding inequality in section 4.

In the present setting with errors, we can also derive a relation which is exactly the same as (2.5) but γ replaced by

$$\tilde{\gamma} := \sum_{j,j'} \epsilon(j \rightarrow j') \text{Tr}[\Pi_j U_{j'}^\dagger \rho_{\text{can}}^{(j')} U_{j'} \Pi_j]. \quad (3.8)$$

Derivation As in (2.7), we use $\sum_i P_i^{(0)} = 1$ and $\sum_j \Pi_j = 1$ to observe that

$$\begin{aligned} \left\langle e^{\beta(E_i^{(0)} - E_k^{(j')})} \frac{Z_0}{Z_{j'}} \frac{\tilde{p}_3(j')}{\epsilon(j \rightarrow j')} \right\rangle_{\tilde{p}} &= \sum_{i,j,j',k} \text{Tr}[P_k^{(j')} U_{j'} \Pi_j U P_i^{(0)} U^\dagger \Pi_j U_{j'}^\dagger P_k^{(j')}] \frac{e^{-\beta E_k^{(j')}}}{Z_{j'}} \tilde{p}_3(j') \\ &= \sum_{j,j',k} \text{Tr}[P_k^{(j')} U_{j'} \Pi_j U_{j'}^\dagger P_k^{(j')}] \frac{e^{-\beta E_k^{(j')}}}{Z_{j'}} \tilde{p}_3(j') \\ &= \sum_{j',k} \text{Tr}[P_k^{(j')}] \frac{e^{-\beta E_k^{(j')}}}{Z_{j'}} \tilde{p}_3(j') = \sum_{j'} \tilde{p}_3(j') = 1, \end{aligned} \quad (3.9)$$

where we noted $\text{Tr}[P_k^{(j')}] = 1$. By recalling the definition (3.5), we get (3.7).

The derivation of (2.5) with γ replaced by (3.8) is essentially the same as that in section 2, and is omitted.

⁹ We note, however, that our equality (3.7) is valid as it is when we take into account quantum mechanical errors by measurement operators M_j which satisfy a certain condition. See section 4.

Error-free limit Consider the limit where $\epsilon(j \rightarrow j')$ tends to $\delta_{j,j'}$, where one recovers the error-free setting of section 2. The (unaverage) mutual information (3.5) becomes $I_{j,j'} \rightarrow \delta_{j,j'} S_j$ with $S_j = -\log \tilde{p}_2(j) = -\log \tilde{p}_3(j)$. Thus the averaged quantity (3.6) becomes $\langle I_{j,j'} \rangle_{\tilde{p}} \rightarrow \langle S_j \rangle_p = -\sum_j \tilde{p}_2(j) \log \tilde{p}_2(j)$, which is the Shannon entropy.

Then one might be tempted to conjecture the validity of a relation corresponding to (3.7) for the error-free setting of section 2, namely,

$$\left\langle \exp \left[\beta \{ W_{i,j,k} - (F^{(0)} - F^{(j)}) \} - S_j \right] \right\rangle_p \stackrel{?}{=} 1. \quad (3.10)$$

Unfortunately, a careful look at the above derivation reveals that this is not valid in general¹⁰. In (3.9), we are making use of the relation $\epsilon(j \rightarrow j')/\epsilon(j \rightarrow j') = 1$, which is not true in the error-free limit. In fact one can easily check that the conjectured (3.10) does not hold in the two-level system considered at the end of section 2.

4 Further observations and discussions

Corresponding inequalities As is always the case, one can use Jensen's inequality $\langle e^f \rangle \geq e^{\langle f \rangle}$ to derive inequalities from the equalities that we have shown. From (2.5), in particular, we get

$$\langle W_{i,j,k} \rangle_p \leq F^{(0)} - \sum_j p_2(j) F^{(j)} + \frac{1}{\beta} \log \gamma, \quad (4.1)$$

where $p_2(j) := \sum_{i,k} p(i \rightarrow j \rightarrow k) = \text{Tr}[\Pi_j U \rho_{\text{can}}^{(0)} U^\dagger \Pi_j]$ is the probability that one gets an outcome j in the intermediate measurement. Note that the left-hand side is written as

$$\langle W_{i,j,k} \rangle_p = \text{Tr}[H^{(0)} \rho_{\text{can}}^{(0)}] - \sum_j p_2(j) \text{Tr}[H^{(j)} \rho_{\text{fin}}^{(j)}], \quad (4.2)$$

where

$$\rho_{\text{fin}}^{(j)} := \frac{U_j \Pi_j U \rho_{\text{can}}^{(0)} U^\dagger \Pi_j U_j^\dagger}{\text{Tr}[\Pi_j U \rho_{\text{can}}^{(0)} U^\dagger \Pi_j]} \quad (4.3)$$

is the final state of the system when the outcome of the intermediate measurement is j . It should be remarked that (4.2) only involves standard quantum mechanical expectation values in the initial and the final states. This is in contrast with the left-hand side of the equality (2.5), which is defined only in a rather artificial setting where one measures the energy both in the initial and the final moments. Since (4.2) is the expectation value of the total work done by the system¹¹, the inequality (4.1) can be interpreted as a generalization of the second law of thermodynamics. This is distinct from the second law derived by Sagawa and Ueda in [5], which is valid for quantum systems with feedback.

¹⁰ It is a tacit assumption in [6] that $P[y|\Gamma_m]$ is always nonvanishing. Although the example of the Szilard engine considered in [6] fails to satisfy the assumption, the relation (4) happens to be valid for certain reasons specific to the model. We thank Takahiro Sagawa for clarifying this point.

¹¹ As we have noted before this includes work exchanged during the measurement.

Obliviously one can replace $e^{\beta(E_i^{(0)} - E_k^{(j)})}$ in the left-hand side of (2.7) by $e^{\beta E_i^{(0)} - \beta_k E_k^{(j)}}$ with an arbitrary β_1, \dots, β_n to get an exact relation. This relation again yields an inequality for $\text{Tr}[H^{(0)} \rho_{\text{can}}^{(0)}]$ and $\text{Tr}[H^{(j)} \rho_{\text{fin}}^{(j)}]$, which is more general than (4.1). It would be interesting to see whether one can get stronger inequalities than (4.1) by choosing optimal β_j which reflect the nature of the operation and the feedback.

From the equality (3.7), one gets another second law

$$\langle W_{i,j',k} \rangle_{\tilde{p}} \leq F^{(0)} - \sum_{j'} \tilde{p}_3(j') F^{(j')} + \frac{1}{\beta} \langle I_{j,j'} \rangle_{\tilde{p}}, \quad (4.4)$$

where

$$\langle W_{i,j',k} \rangle_{\tilde{p}} = \text{Tr}[H^{(0)} \rho_{\text{can}}^{(0)}] - \sum_{j,j'} \tilde{p}_{2,3}(j, j') \text{Tr}[H^{(j')} \tilde{\rho}_{\text{fin}}^{(j,j')}] \quad (4.5)$$

is again the expectation value of the total work, where we wrote

$$\tilde{\rho}_{\text{fin}}^{(j,j')} := \frac{U_{j'} \Pi_j U \rho_{\text{can}}^{(0)} U^\dagger \Pi_j U_{j'}^\dagger}{\text{Tr}[\Pi_j U \rho_{\text{can}}^{(0)} U^\dagger \Pi_j]}. \quad (4.6)$$

The inequality (4.4) looks identical to the second law for quantum systems with feedback derived by Sagawa and Ueda [5], but their inequality contains the QC-mutual information rather than the classical mutual information $\langle I_{j,j'} \rangle_{\tilde{p}}$. The two inequalities are indeed different relations¹². We also note that one can safely take the error-free limit in the inequality (4.4) since nothing like 0/0 is encountered here.

General measurements Let us remark that the same derivation as in section 2 works if one replaces the projections Π_1, \dots, Π_n with general measurement operators M_1, \dots, M_n which satisfy $\sum_j M_j^\dagger M_j = 1$. One gets the Jarzynski-Sagawa-Ueda relation (2.5) and the corresponding inequality (4.1) with γ replaced by

$$\gamma := \sum_j \text{Tr}[M_j^\dagger U_j^\dagger \rho_{\text{can}}^{(j)} U_j M_j]. \quad (4.7)$$

Note that the summand is not the probability of observing M_j in the state $U_j^\dagger \rho_{\text{can}}^{(j)} U_j$ unless M_j is Hermitian, since in (4.7) we have M_j^\dagger instead of M_j . It is expected that M_j^\dagger corresponds to a suitable “time-reversed” measurement. See, for example, [13].

The relation (3.7), on the other hand, can be derived (as it is) only when one has $\sum_j M_j M_j^\dagger = 1$. This condition is valid when all M_j are Hermitian, or (more generally) when $[M_j, M_j^\dagger] = 0$ for all j , but is not valid in general.

¹² If one applies the result of [5] to our setting, one gets an inequality which takes into account “quantum mechanical errors” but is independent of the classical error probability $\epsilon(j \rightarrow j')$ [12]. On the other hand our inequality (which, as we explain below, is valid as it is if the measurement is described by operators M_j which satisfy a certain condition) is apparently insensitive to quantum mechanical errors.

Quantum Jarzynski relation with measurement Consider the error-free setting of section 2, and further suppose that the agent makes a measurement, but do not perform any feedback. This means that we take $U_j = U'$, $H^{(j)} = H'$, and $\rho_{\text{can}}^{(j)} = \rho'_{\text{can}}$ for any outcome $j = 1, \dots, n$.

In this case we see from (2.6) that

$$\gamma = \sum_j \text{Tr}[\Pi_j (U')^\dagger \rho'_{\text{can}} U' \Pi_j] = \text{Tr}[(U')^\dagger \rho'_{\text{can}} U'] = \text{Tr}[\rho'_{\text{can}}] = 1, \quad (4.8)$$

and hence our main result (2.5) reduces to the quantum version of Jarzynski relation [10]. Since a measurement process generally disturbs a quantum state, it is a nontrivial fact that the quantum Jarzynski relation gets no modifications here. This is closely related to the extension of fluctuation theorem in [8].

When projective measurement is replaced by more general measurement as above, we have $\gamma = 1$ only when the condition $\sum_j M_j M_j^\dagger = 1$ is valid.

If one takes the setting of section 3 with errors, and assumes that $\epsilon(j \rightarrow j')$ is independent of j , one immediately gets $\tilde{\gamma} = 1$ from (3.8). Thus we also have the standard Jarzynski relation in this case. This is natural since one has $p_3(j') = \epsilon(j \rightarrow j')$ and hence $I_{j,j'} = 0$, i.e., the agent is making no use of information.

On “clockwork demon” We have here assumed the existence of an external intelligent agent (demon) who performs feedback taking into account the outcome of the intermediate measurement. It is also interesting to consider the possibility of designing an external quantum mechanical system, a “clockwork demon”, which is programmed to perform the exact feedback we want¹³.

The Hamiltonian of the whole system is written as $H_{\text{tot}}(t) = H_{\text{system}}(t) + H_{\text{demon}}(t) + H_{\text{int}}(t)$, and its time-dependence is fixed in advance. The interaction Hamiltonian $H_{\text{int}}(t)$ is vanishing in the initial and the final moments. Initially both the system and the demon are in their (canonical) equilibrium states¹⁴ with a common β . Then the demon Hamiltonian $H_{\text{demon}}(t)$ and the interaction Hamiltonian $H_{\text{int}}(t)$ are changed to let the demon interact with the system as designed. The actual design of such an autonomous feedback system is far from trivial. It is indeed a nontrivial question whether one can realize an arbitrary feedback scenario in this manner, but we won’t go into details here.

It is crucial that the original Jarzynski relation without feedback [4, 10] applies to this situation. Since the system and the demon are decoupled in the initial and the final moments, the relation reads

$$\left\langle \exp \left[\beta \{ W - (F_{\text{system}}^{\text{init}} - F_{\text{system}}^{\text{fin}}) \} \right] \right\rangle = \exp [\beta (F_{\text{demon}}^{\text{init}} - F_{\text{demon}}^{\text{fin}})], \quad (4.9)$$

¹³ The following discussion of course applies to the classical setting (as in [6]) as well. We also stress that this is quite a standard thought, which has appeared in many different contexts.

¹⁴ In many cases, it is more natural to assume that the demon is initially in a quasi-equilibrium state. More precisely, one assumes that the state of demon is initially restricted to a certain subspace of the whole Hilbert space, and distributed according to the canonical distribution within the subspace. Then the original Jarzynski relation does not apply, and the following discussion is not valid as it is.

where W is the difference between the initial total energy and the final total energy. We here make a nontrivial and crucial assumption that the clockwork demon is designed in such a way that it does not exchange appreciable work with the system or with the agent who changes the Hamiltonian¹⁵. Then the work W is interpreted as essentially the work done by the system to the external agent, and the left-hand side of (4.9) can be identified with that of (2.5). Comparing (4.9) with (2.5), one sees that γ , in this case, is directly related to the free energy difference of the clockwork demon.

Let us stress, however, that this consideration does not diminish the significance of Sagawa and Ueda's generalization nor reduce it to the original Jarzynski relation. The power of the generalized relation appears, for example, in the compact representation (2.6) of γ , which does not only suggest a clear interpretation but also makes it possible to measure γ experimentally. One can say that, by considering an idealized setting with feedback, Sagawa and Ueda were able to sort, in quite a neat manner, complicated exchange of heat and work into the part which is directly related to the feedback and the part intrinsic to the system.

We wish to thank Takahiro Sagawa for indispensable discussions and comments, Jordan Horowitz and Akira Shimizu for useful discussions.

References

- [1] D. J. Evans, E. G. D. Cohen, and G. P. Morriss, Probability of second law violations in shearing steady states, *Phys. Rev. Lett.* **71**, 2401–2404 (1993).
- [2] G. Gallavotti and E. G. D. Cohen, Dynamical Ensembles in Nonequilibrium Statistical Mechanics, *Phys. Rev. Lett.* **74**, 2694–2697 (1995), [chao-dyn/9410007](#).
- [3] G. E. Crooks, Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences, *Phys. Rev.* **E60**, 2721–2726 (1999), [cond-mat/9901352](#).
- [4] C. Jarzynski, Nonequilibrium Equality for Free Energy Differences, *Phys. Rev. Lett.* **78**, 2690 (1997), [cond-mat/9610209](#).
- [5] T. Sagawa and M. Ueda, Second Law of Thermodynamics with Discrete Quantum Feedback Control, *Phys. Rev. Lett.* **100**, 80403 (2008), [arXiv:0710.0956](#).
- [6] T. Sagawa and M. Ueda, Generalized Jarzynski Equality under Nonequilibrium Feedback Control, *Phys. Rev. Lett.* **104**, 90602 (2010), [arXiv:0907.4914](#).
- [7] S. Toyabe, T. Sagawa, M. Ueda, E. Muneyuki, and M. Sano, Experimental demonstration of information-to-energy conservation and validation of the generalized Jarzynski equality, *Nature Phys.* **6**, 988–992 (2010), [arXiv:1009.5287](#).

¹⁵ We still do not know in which case this (rather tricky) assumption is realized.

- [8] M. Campisi, P. Talkner, P. Hänggi, Fluctuation Theorems for Continuously Monitored Quantum Fluxes, *Phys. Rev. Lett.* **105**, 140601 (2010), [arXiv:1006.1542](#).
- [9] J. M. Horowitz and S. Vaikuntanathan, Nonequilibrium Detailed Fluctuation Theorem for Repeated Discrete Feedback, preprint (2010), [arXiv:1011.4273](#).
- [10] H. Tasaki, Jarzynski relations for quantum systems and some applications, unpublished note (2000), [cond-mat/0009244](#).
- [11] J. Kurchan, A quantum fluctuation theorem, preprint (2000), [cond-mat/0007360](#).
- [12] T. Sagawa, private communication.
- [13] H. Terashima and M. Ueda, Hermitian conjugate measurement, *Phys. Rev.* **A81**, 1094–1622 (2010), [arXiv:0709.1210](#).